The Strategy Unit.

How will we know if Integrated Care Systems reduce demand for urgent care?

Establishing fair benchmark levels for the blended payment system

January 2020

Prepared by:
Andrew Jones

NHS
Midlands and Lancashire
Commissioning Support Unit

Document control

Document Title	Establishing fair benchmark levels for the blended payment system
Job No	
Prepared by	Andrew Jones, Steven Wyatt, Paul Seamer
Checked by	Steven Wyatt
Date	January 2020

Contents

Foreword

"Each system is perfectly designed to get the results it gets." [1]

For most NHS services, healthcare commissioners pay providers according to the rules and prices of the National Tariff Payment System (NTPS). Until April 2019, the NTPS operated on a fee-for-service basis. At face value, the fee-for-service model appears to offer providers an incentive to increase supply and therefore sets the financial sustainability of providers against that of the health system as a whole.

It is believed that such tensions might be resolved with the introduction of risk-and-reward-sharing, or "blended payment", schemes. This alternative payment model encourages the provider to moderate growth in activity by assigning them a share of the annual savings or the cost over-runs. The risk-reward sharing model is currently seen as the most appropriate way to distribute resources in the healthcare system and, as a consequence, the NTPS has recently adopted blended payments for emergency activity.

Central to the blended system is the recommendation that commissioners and providers reach agreement on "planned" activity levels (the future activity that might be expected under normal circumstances). The provider's subsequent performance is measured relative to these levels, and rewards or penalties allocated. It is therefore vital that these levels be fair and credible, and that the methods used to create them be authoritative and transparent. Moreover, these levels must be calibrated to support the objectives of the national healthcare system.

Yet, official documents currently offer little detail on these crucial planned levels, and no firm guidance on how to produce them. This is problematic since falling back on conventional forecasting methods in this context may lead to the unfair allocation of millions of pounds worth of incentives and a missed opportunity to improve commissioner-provider relations.

This paper has three objectives:

i. To illustrate the workings of the blended payment system.

ii. To demonstrate that inappropriate modelling of "planned" activity levels could divert tens of millions of pounds away from the emergency care system.

iii. To pinpoint the reasons why conventional forecasting approaches are unsuitable in this context, and to suggest alternatives.

It may be tempting to consider the blended payment system as a technical tool for determining the allocation of resources between commissioners and providers - something that can be left to analysts and finance specialists to negotiate. But the implications of the

blended payment system are far reaching: Decisions about planned activity levels will determine the total funding envelope for urgent care within a system and will influence the behaviour of healthcare providers and the services they deliver to patients. This is an unusual situation where senior managers in commissioner and provider organisations must engage in the detail, however esoteric it may seem, to ensure both the financial sustainability of their organisations and the quality and accessibility of services for the populations they serve.

[1] Paul B. Bataldan, MD, Senior Fellow, Institute for Healthcare Improvement.

1. The Blended Payment System

1.1 Why move to a blended payment system?

An NHS commissioner will buy services from one or more providers through a series of different contracts[2]. For the majority of NHS-funded services, including emergency admissions, A&E visits, and same-day activity, these contracts are underpinned by the National Tariff Payment System (NTPS).

Broadly[3], there are three models by which the NTPS could compensate a provider for the activity it carries out:

1. A fee-for-service contract, under which the provider is paid a fixed price for each unit of activity performed.
2. A capitated (fee-per-patient) contract, where the total paid to the provider is based on the population covered by the provider.
3. A risk-reward sharing (or blended payment) contract, under which both the commissioner and provider(s) may share in savings or cost over-runs, contingent on performance against a benchmark activity level.

Each of these models assigns the commissioner and provider a different share of the risk associated with cost growth. In a fee-for-service model, the provider has an incentive to increase supply[4] thus the commissioner faces the risk. With a capitated contract, the commissioner fixes the budget for the year and the risk of unexpected growth lies with the provider(s). Blended payment (or risk and reward sharing) contracts offer a compromise between the two extremes.

For the 2019/20 financial year the NTPS for emergency care moved from a fee-for service arrangement to a blended payment system. The blended system encourages the provider to moderate activity growth by providing financial incentives for effective demand management. A well-designed blended payment system may also lead to greater collaboration between provider and commissioner, more effective planning and forecasting, and a more innovative healthcare system[5].

1.2 How do blended payments work?

The blended payment from commissioner to provider consists of two elements:

1. **A fixed payment.** Before a new financial year begins, providers and commissioners should agree on estimated (or "planned") activity levels for following 12 months. This activity level is then multiplied by HRG prices to determine the fixed payment element.
2. **A volume-based element.** At the end of the financial year, the difference between the planned and the observed activity levels will dictate the volume-based payment. If the observed activity level is greater than planned, providers receive the fixed payment *plus* 20% of the difference between the final priced activity and the planned value. If the observed level is lower than planned, providers receive the fixed payment *minus* 20% of the difference between the planned and the final priced levels.

Figure 1, overleaf, illustrates how these two elements ultimately determine financial benefits, or penalties, for providers and commissioners. If observed activity exceeds the planned levels, each is required to pay a share of cost over-runs. However, if observed activity levels are lower than planned, both the provider and commissioner will gain financially: the provider is paid more than the priced activity and the commissioner pays less than the expected amount.

Under the blended payment system, commissioners (purchasers) and providers will continue to have competing financial interests: Commissioners will wish to set the planned activity level as low as possible[6] (to minimise the fixed payment to the provider) since they will pay less than the total costs if observed activity is greater than planned (Figure 1, top). By contrast, the provider will wish to set the planned level relatively high to maximise the potential for (and level of) savings, and minimise the risk of penalties (Figure 1, bottom).

Yet, despite these tensions, those designing the blended payment system highlight the potential for increased understanding and collaboration between provider and commissioner due to the shared planning process. Trials have suggested that the blended approach to payments can move the focus away from "challenging [the] activity recording from a financial perspective, to improving the quality of coding to improve clinical decision making." [7]

Figure **1**. An illustration of blended payments in two scenarios. Where the observed activity exceeds the planned level (top), the commissioner pays some of the cost over-runs (the volume-based element), but the provider will incur an overall loss. In a situation where the actual activity is less than the planned level (bottom) the commissioner pays the provider the fixed payment minus the volume-based element. The provider is left with an overall surplus. Note: We have included the effect of marginal rates and assumed that these are symmetric (reduction in cost when activity is reduced by one unit is equal to the increase in cost associated with delivering one additional unit) and fixed at 30%.

1.3 Do we require planned activity or benchmark activity levels?

As the reference level from which rewards and penalties are calculated, "planned" activity levels will ultimately determine the allocation of millions of pounds and may have considerable influence on relations between commissioners and providers. Yet, information on how to produce these markers is scant[8]. According to the document, Guidance on Blended Payment for Emergency Care[9],

"[Planned activity levels] should reflect the effects of demographic pressures as well as [provide a] realistic assessment of the impact of system efforts to reduce demand."

Now, we would commonly take "planned levels" to mean, "our best attempt to predict activity levels". Yet, in the context of the blended payment system, this interpretation will likely lead to undesirable outcomes.

For example, the guidance[10] suggests that, in order to produce planned levels, we first establish an activity baseline from historic levels. To this, we would apply adjustments, post hoc, to account for factors such as demographic change and demand management initiatives (e.g. QIPP[11] schemes). However, we must treat the demand management component with care, for if we wholly incorporate the estimated effect of these strategies (a reduction in activity) into the planned level, we leave no room for providers to make additional reductions[12] and penalise all providers whose demand management strategies fell short of the expected reductions. The result is tantamount to a one-sided risk-reward sharing model in which the full risk of activity growth lies with the provider.

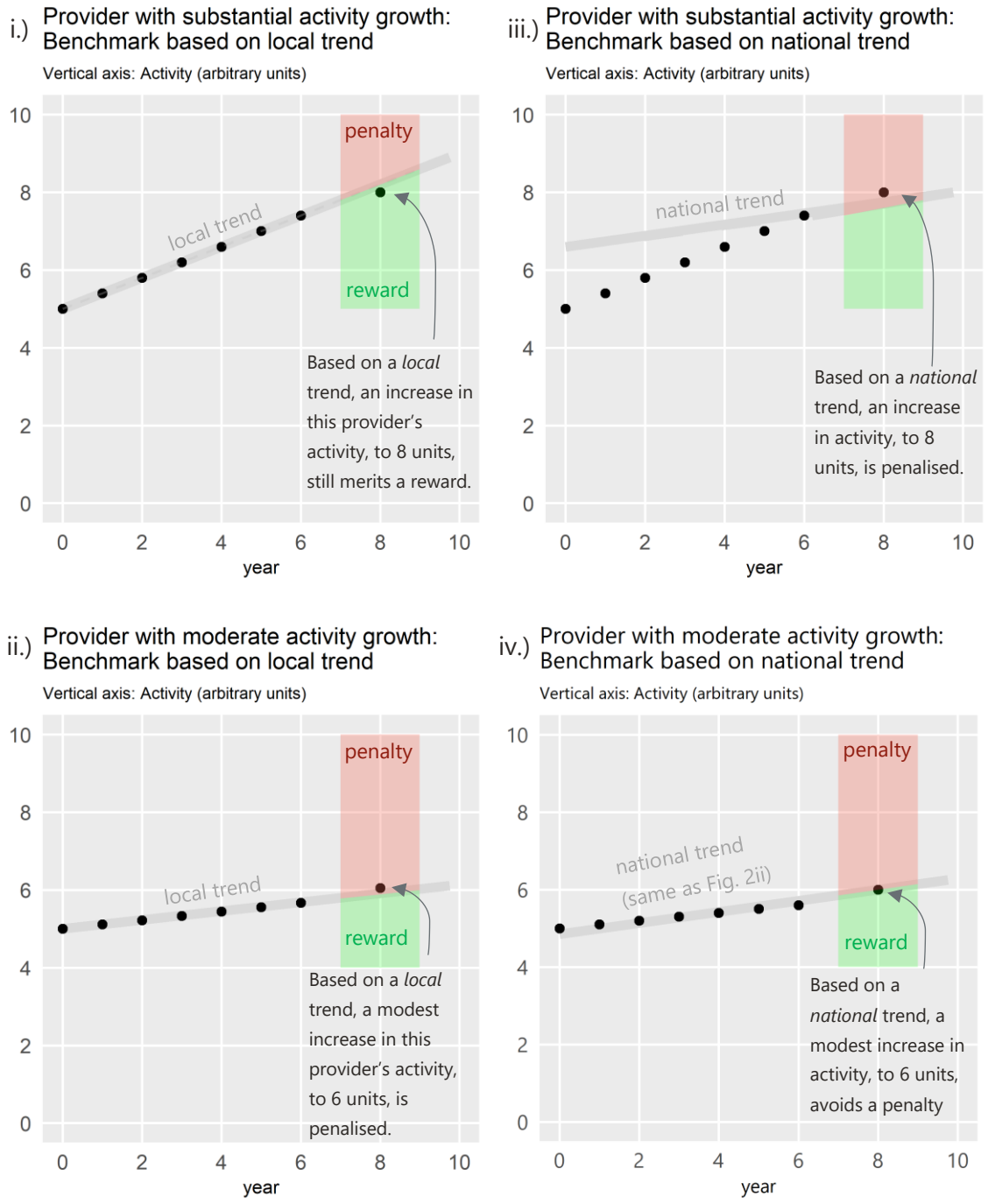
An alternative approach is to capture the continuous efforts of healthcare systems to manage demand (along with other non-demographic factors[13]) in a trend component. But here, again, much will depend on our interpretation of "planned level".

To illustrate, let us imagine we have been asked to produce planned activity levels for Provider X, and that, in addition to modelling demographic changes, we have decided to capture the effect of demand management strategies in a trend component. If we understand the planned level to be "our best attempt to predict activity levels", we would likely regard Provider X's own record of past performance (the local trend) as the most appropriate guide to future performance.

Now, if Provider X had a history of substantial activity growth then – all else being equal - the local trend method will predict substantial activity growth in the future. Thus, a planned level based on X's local trend would accommodate immoderate activity growth and may ultimately result in rewards for this behaviour (**Figure 2i**).

On the other hand, if Provider X had successfully *moderated* activity in recent years, a planned level based on a local trend would require the provider to match or exceed past levels of success to avoid penalty (**Figure 2ii**).

Figure **2**. Having controlled for demographics: **i)** An inefficient provider could be rewarded for immoderate activity growth if activity is modelled using a local trend. **ii)** By contrast, if we employ a national trend the provider must do better than the historic national average to earn a reward. **iii)** A provider who had successfully moderated activity growth in recent years could be penalised for a modest increase if activity is modelled using a local trend. **iv)** Using a national trend, there is a penalty only if the rate of activity growth was greater than the historic national average.



Rather than following this local, "purely predictive" forecasting approach, we might consider replacing each provider's unique trend (the local trend) with the national trend. In this case, we would reward providers if the rate of activity growth was lower than the historic national average (**Figure 2iii** and **Figure 2iv**). We suggest that, for a system operating on a national level, it is more appropriate to reward a provider in relation to their improvement relative to the average, than to offer incentives based on a provider's own past performance.

In this section, we have seen how various interpretations of "planned level" might influence payment outcomes. We suggest that a first step[14] in overcoming the problems associated with interpretation would be more precise nomenclature. In this regard, we believe "benchmark activity level" more closely reflects the objectives of the modelling process in this context, and we will be using this term in the analysis that follows.

[2] With the increasing integration of healthcare services, there are plans to provide a single commissioning contract to support the integrated service model. The Draft Integrated Care Provider (ICP) Contract (NHS England, 2018) is designed to increase collaboration across the system, allowing better coordination of care and a broader focus on the needs of the individual. Such contracts could last for up to 10 years, offering stability to both providers and commissioners, bringing about a shared focus on long-term outcomes. The proposed Integrated Care Provider (ICP) Contract would be partly underpinned by the blended payment system.

[3] If we leave to one side the idea of bundles and block contracts

[4] This is due to marginal costs (i.e. the cost of providing one additional unit of activity tends to diminish as total activity increases).

[5] For a broader discussion, see *Risk and Reward Sharing for NHS Integrated Care Systems*, Strategy Unit (2018)

[6] Assuming the desire to moderate growth.

[7] *Guidance on blended payment for emergency care*. NHS England and NHS Improvement (March 2019)

[8] Unlike the highly detailed process for setting the equivalent "counterfactual" activity levels for risk-reward sharing schemes in the United States.

[9] *Guidance on blended payment for emergency care*. NHS England and NHS Improvement (March 2019)

[10] Ibid.

[11] Quality, Innovation, Productivity and Prevention

[12] Although it would still be possible for activity to be lower than planned, this would likely be due to factors beyond the providers control (e.g. technological innovation) and rewards would therefore be unmerited.

[13] E.g. changes to clinical practice or standards, or the introduction of new technology

[14] Ultimately new nomenclature should be underpinned by a detailed set of methods.

2. Comparing Three Approaches to Modelling Benchmark Activity Levels for Emergency Admissions

Having discussed theoretical differences between approaches to modelling benchmark activity levels, we now look at concrete examples based on cases outlined in Section 1.3. We also examine the financial implications of each approach.

We reiterate that the NTPS (and hence the blended payment system) currently operates at the Clinical Commissioning Group (CCG) and provider level. Whilst CCGs serve clearly defined populations, it is difficult to define the population served by a single provider (with the added complication that market share changes over time). **We have therefore chosen to present results at the Integrated Care System (ICS) / Sustainability and Transformation Partnership (STP) level.** This way, the flow of patients to providers outside a given footprint should be a small proportion of the total admissions in that footprint.

In any case, the principles we demonstrate in this simplified environment should be applicable no matter the level at which contracting methods operate.

2.1 Data

We utilised twelve years of English admissions data (2005 to 2016) from Hospital Episode Statistics (HES). We counted all emergency activity that was coded as the first episode in a spell, and removed admissions falling under the "Well Babies" treatment specialty.

To replicate the lag that occurs between activity and the availability of the HES data that represents that activity, we used nine years of our data extract (2005-2013) to train models and the final year of data (2016) to validate our model. For the training split, we established population denominators for each stratum using ONS population *estimates* for local authorities in England. For the validation split, we utilised (2014 based) ONS local authority population *projections* as denominators, since we would ultimately require models to forecast emergency activity several years beyond the available admissions data. Finally, as the population projections group individuals over the age of 90 into a single stratum, we were obliged to do the same with all our admissions data.

We derived rates of admission by year, age, gender and council with adult social services responsibility (council area). We followed standard practice and removed two council areas with particularly small resident populations (City of London, and Isles of Scilly).

2.2 Models

As the blended payment guidance suggests, we might reasonably expect a model that predicts healthcare activity to account for demographic pressures (the changing age and sex structure of population) and "system efforts to reduce demand" (which, as we discussed in Section 1.3, might adequately be captured with a trend component).

Based on these basic determinants of healthcare utilisation, we will compare three possible approaches to setting planned activity levels:

1. **The common** (NHS) **approach *(local demographics only)***
   This is a simple cross-sectional model which is widely used in the NHS when forecasting activity[15] (details in Box 1). The model will predict activity levels based solely on demographic pressures (the change in demographic structure of the population served). There is no accounting for non-demographic factors[16].
2. **A "local trend" approach *(local demographics + local trend)***
   This is a more powerful alternative to the common approach. We used a generalised additive model (GAM). The model used here will make predictions of activity levels based on changes in the local demographic structure and the local (linear) trend[17].
3. **A "national trend" approach *(local demographics + national trend)***
   This is also a GAM, but predictions are based on changes in the local demographic structure as well as the national (linear) trend. As suggested in Section 1.3, we believe this model may better support the objectives of the blended payment system.

Box 1: The most common approach to estimating the impact of demographic growth on a form of healthcare (e.g. emergency admissions, or primary care consultations) consists of the following steps:

1. Count the number of units of healthcare delivered in the baseline year, by gender and age group.
2. Using sub-national population projections, calculate the growth in population, by age and gender, from the baseline year to the future year of interest.
3. Multiply the baseline units of healthcare by the growth in population for each gender and age group.

In the appendix, we provide details of an analysis which compares the prediction accuracy of the common (NHS) approach to modelling activity with a simple generalised additive model (GAM).

We found that a GAM, fitted to each council area, consistently produced more accurate predictions than the common approach. The basic formulation was;

$$\text{rate of admission} = B_0 + f(\text{age group X gender}) + B_1 * \text{gender} + \varepsilon, \quad \text{(Eq. 1)}$$

where B_0 and B_1 are model parameters, f is a smooth function of the age variable (interacting with gender), and ε is a mean-zero random error term.

Having established confidence in the GAM approach, we added a trend component to Eq. 1 (following our interpretation of the guidance in the blended payments document);

$$\begin{aligned}\text{rate of admission} = {} & B_0 + f(\text{age group X gender}) \\ & + B_1 * (\text{gender}) \\ & + B_2 * (\text{year}) + \varepsilon . \end{aligned} \quad \text{(Eq. 2)}$$

Here, we have assumed that the relationship between year and rate of admission is linear[18] as this simplifies the forecasting process. This is our "local trend" approach.

For the national average[19] trend, we used a single principle model to cover all council areas, and removed the interaction between council area and year;

$$\begin{aligned}\text{rate of admission} = {} & B_0 + f(\text{age group X gender X council area}) \\ & + B_1 * (\text{gender X council area}) \\ & + B_2 * \text{year} + \varepsilon . \end{aligned} \quad \text{(Eq. 3)}$$

Predictions from the model were produced at local authority level and aggregated to ICS/ STP level.

[15] As highlighted in NHS Five Year Forward View Recap briefing for the Health Select Committee on technical modelling and scenarios. NHSE (2016)

[16] The impact of non-demographic factors would have to be addressed separately

[17] The trend component is a catch-all for non-demographic factors (changes to clinical standards, demand management strategies, technology, etc). The trend in the "purely predictive" GAM is influenced by local non-demographic factors while the trend for the "benchmarking" GAM is influenced by non-demographic factors on a national level.

[18] This assumption appears reasonable based on graphical analysis.

[19] These are "council area" averages

3. Results

As noted at the beginning of Section 2, we have chosen to present results at the ICS/STP level to side-step issues associated with defining provider populations. Each of the 42 ICSs/STPs is composed of several CCGs and is served by a number of providers. **Therefore, where we detail rewards or penalties in the following sections, it should be noted that these payments are shared across the constituent provider organisations.**

In **Figures 4 and 5**, we show the numbers of emergency admissions for six randomly selected ICSs/STPs, for the years 2005 to 2016. In the same figures, we present the 2016 benchmark levels that would have been produced by the common approach, the GAM[20] with a local trend, and the GAM with a national trend

While there is no way to formally assess each model's ability to produce a suitable benchmark, the results, visible in **Figures 4 and 5**, appear to support the theory laid out in Section 1.3. That is, by using the national trend, rather than the local trend or common approach, we are less likely to see unreasonably high (or low) benchmark levels relative to the observed activity.

We see that the common approach consistently underestimates the observed activity since it does not account for non-demographic increases in demand. The two GAMs often produce similar results (as the local trend is frequently comparable to the national trend) but anomalous local trends will occasionally create disparities between the two.

Figure 6 illustrates how reward or penalty payments[21] for the group of providers within each ICS/STP would have differed depending on the approach taken to produce the benchmark level. We see considerable variation around the size of payments, and, in some cases, the models disagree as to whether a reward or penalty is appropriate. As noted, the common approach tends to underestimate the observed activity and therefore sets consistently low benchmark activity levels. In our hypothetical 2016 scenario, providers in thirty-six of the forty-two ICSs/STPs would have exceeded the benchmark level (and faced a penalty) were the common approach used. Examining the two GAMs, the local trend approach tends to allocate larger rewards than the national trend approach, though neither consistently allocates greater penalties.

As **Figure 6** and **Table 1** illustrate, all three approaches assign more penalties than rewards for the year 2016. The skew is perhaps to be expected in this hypothetical scenario (payments in 2016 operated under a fee-for-service system which may have incentivised activity). Influenced, instead, by the incentives of the new payment system, we might expect fewer providers to incur large penalties (and fewer penalties overall[22]). Yet, regardless of the system in place, we would expect to see non-trivial differences between the rewards and penalties allocated by these three modelling approaches.

Figure **4**. A comparison of STP trends and 2016 "benchmark" admissions produced by three types of model. Taking North West London (top left panel) as an example, we see admissions have increased steadily over the period. The grey point shows the observed admission levels for 2016. This observed level is higher than that predicted by the common approach (turquoise point). If the common approach were used to set the planned activity levels for North West London, we would see penalties applied. By contrast, we see the observed level is below both the dark blue point (local trend approach) and dark red point (national trend approach). If we were to use either of these as benchmarks, the providers would receive a reward (the purely predictive approach would, however, allocate a greater reward). It bears repeating that the national trend approach does not attempt to forecast the number of admissions but produces a benchmark for activity based on population structure and the historic national trend. The difference between benchmarking and observed levels should reflect improvements in efficiency relative to the national rate. Note that in all cases the vertical axis has been truncated to highlight the different results.

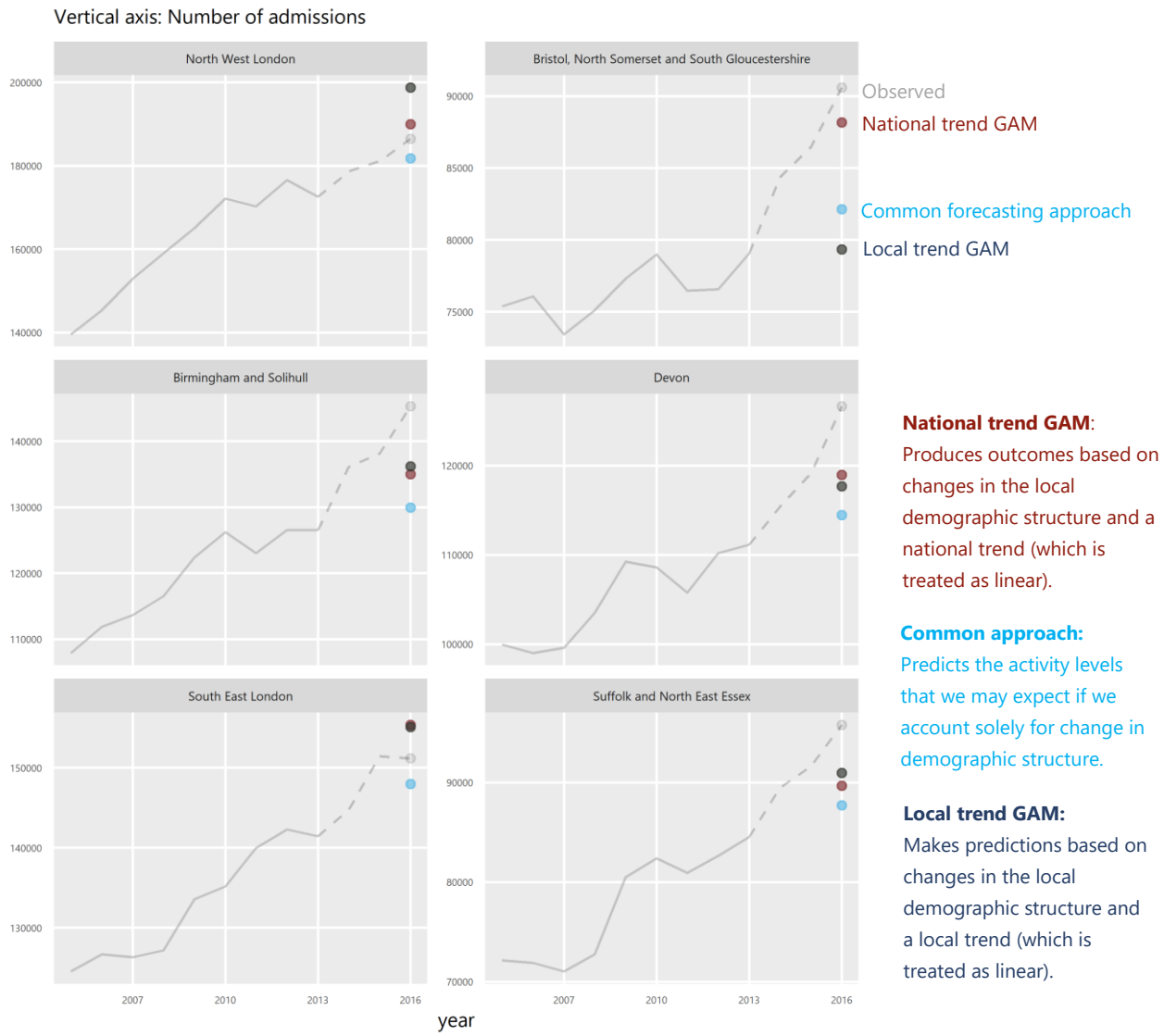
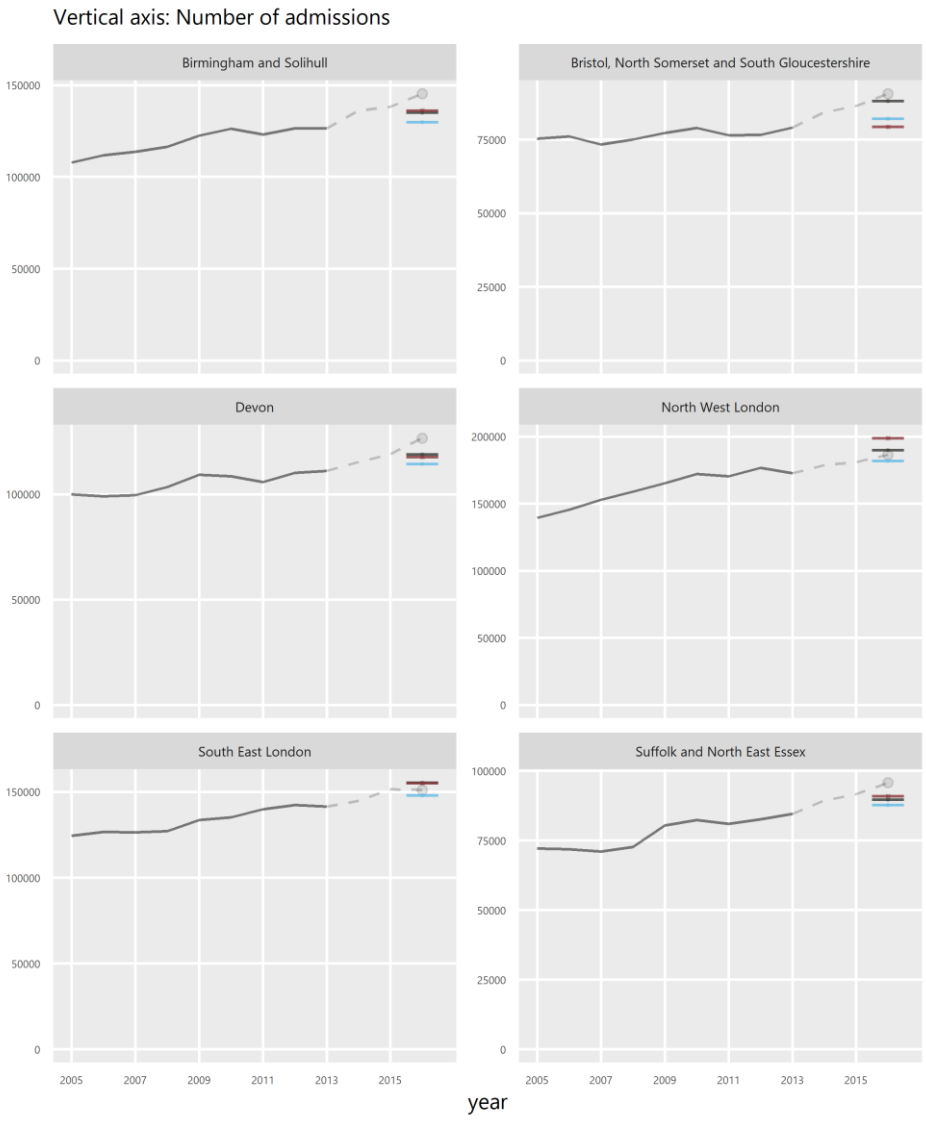


Figure **5**. We present the same data as in Figure 4 but, in this case, we do not truncate the vertical axis. The differences between benchmark levels produced by these approaches are between 1 and 10 % of the total number of admissions.

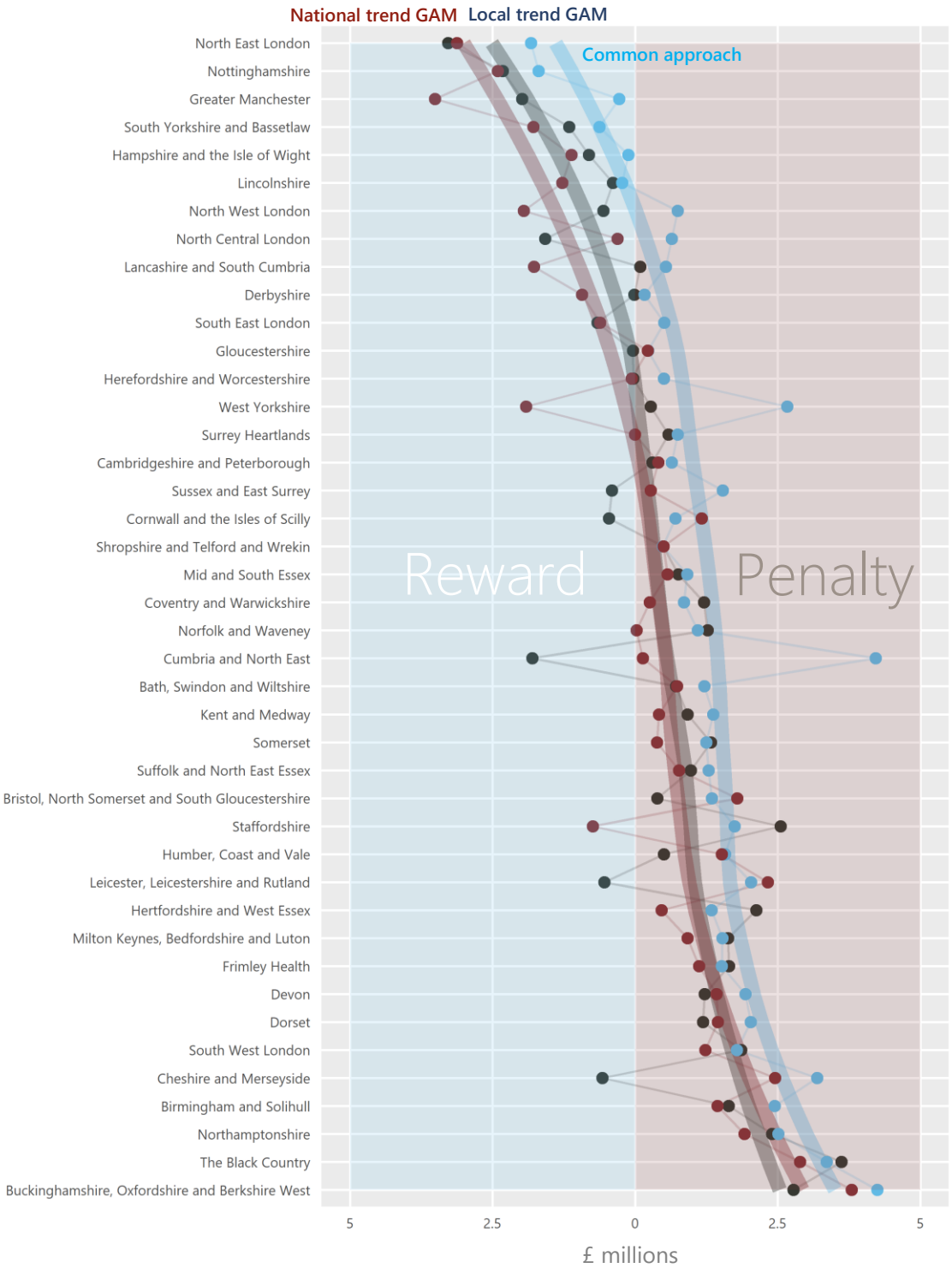


National trend GAM: Produces outcomes based on changes in the local demographic structure and a national trend (which is treated as linear).

Common approach: Predicts the activity levels that we may expect if we account solely for change in demographic structure.

Local trend GAM: Makes predictions based on changes in the local demographic structure and a local trend (which is treated as linear).

Figure **6**. A comparison of hypothetical rewards and penalties for STPs in the year 2016 determined by three different modelling approaches. The total cost figures are based on the average unit cost for a non-elective inpatient admission in 2016/17, which was £1,590[23].



The possible implications of each approach to the national emergency care system can be seen in **Table 2**. A model that sets benchmark levels unreasonably low (as the common approach does) could remove tens of millions of pounds[24] from the emergency care system. Conversely, an approach that is overly tolerant of activity growth (as the local trend approach may be) will not properly incentivise providers to moderate demand. Once again, these figures relate to emergency admissions, alone.

Table **1**. Total rewards and penalties assigned by the different modelling approaches applied to 2016 activity.

Model	Total rewards (£ millions)	Total penalties (£ millions)
National trend GAM	16.6	32.5
Local trend GAM	21.5	30.6
Common approach	4.8	54.9

Table **2**. Total additional reward payments (at national level) allocated by models (when compared to the national trend GAM approach). This may also be thought of additional payments into emergency care system.

Model	Total additional reward payments compared to national trend GAM (£ millions)
Local trend GAM	6.8
Common approach	-34.3

[20] Both GAMs cover the age groups from 1 to 89. Supplementary, negative binomial regression models were used for the age groups of 0 and 90. Further details can be found in the appendix.
[21] For emergency admissions
[22] A related point: If we use an historic trend, based on activity in a fee-for-service system (which incentivised activity), we would expect activity levels to be higher before the introduction of the blended system than after it. Thus, in first few years of a new system we would expect rewards for providers to be likely. This is perhaps a reason why, when risk and reward sharing contracts were introduced in the United States, many providers signed up.
[23] *Reference costs 2017/18: highlights, analysis and introduction to the data*. NHS Improvement (2018)
[24] The figures in Table 2 are due to emergency hospital admissions, alone.

4. Discussion

This analysis has highlighted the implications of ambiguous guidance around approaches to setting "planned" activity levels within the blended payment system. Carefully calibrated "planned" levels are crucial to the effective operation of the blended system since they determine the allocation of many millions of pounds of rewards (or penalties) to providers. Without a firm definition of what is required, we will likely see a range of interpretations, unfair rewards, and unnecessary friction between NHS providers and commissioners.

The financial implications of using inappropriate modelling approaches - both to individual providers and the emergency care system as a whole - are non-trivial. Our analysis showed that a group of providers could see swings of up to £6 million in their payments (for emergency admissions, alone), depending on the method used. The overall difference to the emergency care system could be up to £40 million.

Our first suggestion here concerns nomenclature. In the context of the blended payment system, we believe the term "benchmark" (rather than "planned") is a more appropriate descriptor for the level of activity that determines the fixed payment element. This term should invite a greater level of inquiry into what is required, whilst also conveying that there is nuance to the forecasting processes involved. We believe this change should be supported by a precise definition of "benchmark" and perhaps even a standard set of methods for producing one. This would lessen the potential for disputes between commissioners and providers.

Having looked first at the mechanics of the blended system, we went to explore the shortcomings of conventional forecasting approaches when the objective is to produce a "benchmark" level. We suggested how a model designed to produce a suitable benchmark might instead be formulated. For example, an approach that considers local demographics but includes a national trend component would reward providers whose rate of activity growth was lower than the historic national average. Rewards would apply regardless of a provider's overall performance relative to other providers. We will refer to this (the local trend approach) as the "benchmarking approach".

This proposed benchmarking approach is a starting point. This model relies on a historic national trend and may therefore have some unusual characteristics. Firstly, we have assumed that the trend is linear as this simplifies the forecasting process. And while the linear assumption seems to be reasonable when we look backwards in time, we cannot say that this will hold in the future. Second, provider rewards would be more likely in the first years of the new system, since the historic trend would still reflect the old system (which incentivised activity). Furthermore, it would be theoretically possible for all providers to

receive a bonus (or for all to be penalised) if they outperformed (or failed to match) the historic national trend. This last point, however, would likely occur only as a result of great systemic change from one year to the next.

It must again be noted that the benchmarking model is designed to produce counterfactual outcomes and we therefore have no way to directly validate this approach. However, since the GAM models are largely interpretable, we can guess at how they should generally behave. In this way, the local trend GAM becomes a helpful reference for the national trend (benchmarking) GAM. We might also use the former to infer details regarding the fit and suitability of the latter.

4.1 Further work

For certain STPs/ICSs it may be unfair to measure demand management success relative to the national trend. An improved approach might stratify STPs/ICSs according to the healthcare needs and deprivation levels of the populations they serve. These additional factors could be readily incorporated into our benchmarking approach.

Finally, we have accepted a lag of several years between the activity taking place and the availability of the corresponding HES data for modelling purposes. This lag, however, could be minimised with the use of activity data from the National Commissioning Data Repository (NCDR). The historic national trend component would then more relevant to the current situation.

5. Conclusion and Recommendations

Risk-and-reward sharing is currently seen as the most appropriate way to distribute resources across the healthcare system. With the introduction of blended payments, which support the risk-reward sharing model, healthcare commissioners and providers will be required to reach agreement on benchmark activity levels (the future activity levels that might be expected under normal circumstances). These levels will ultimately determine the allocation of millions of pounds of provider rewards and penalties. However, guidance on producing these levels is scant and the little that currently exists could undermine the objectives of the new system. Failure to address this issue may not only lead to the inappropriate distribution of resources across the system; it could result in tens of millions of pounds being diverted away from emergency care.

We offer a number of recommendations for health systems adopting risk and reward payment approaches, such and the blended tariff for emergency care.

1. The term 'benchmark' should be used to describe the level of activity or costs below which rewards are accrued.

2. Benchmarks should take account of projected local demographic change plus the national residual (non-demographic) growth rate.

3. The methods used to set benchmarks and the benchmark levels should be set out in detail, along with a statement of the intended effect of incentives on commissioner(s) and provider(s).

Furthermore, we recommend that additional work is undertaken to develop the methods set out in this paper to incorporate the effects of deprivation on healthcare activity growth.

Appendix

Simple Approaches to Modelling Emergency Admissions: A comparison

The analysis in this appendix compares the common approach to forecasting emergency admissions (detailed in Box 1) with a more sophisticated alternative from the field of statistical regression. We use the same input data for both models.

We sought an alternative to the common approach which might increase the accuracy of predictions while remaining interpretable and versatile. As a result of these requirements, we ruled out some of the more involved machine learning approaches and settled on a form of regression analysis.

Data

We used English admissions data for the years 2013 and 2016 from Hospital Episode Statistics (HES). We counted all emergency activity that was coded as the first episode in a spell, and removed admissions falling under the "Well Babies" treatment specialty.

Models were built on data from the year 2013, while 2016 data was used to compare the results. For 2013, we established population denominators for each stratum using ONS population estimates for or England. For 2016, we utilised (2014 based) ONS population projections to provide denominators, as we ultimately require models to forecast emergency activity several years beyond the available admissions data. Since these population projections group individuals over the age of 90 into a single stratum, we were therefore obliged to do the same with all admissions data.

We derived rates of admissions by age, gender and council with adult social services responsibility (council area). We followed standard practice and removed two council areas with particularly small resident populations (City of London, and Isles of Scilly).

Common approaches to predicting healthcare activity, and their limitations

Within the NHS, the most common method for forecasting (emergency admissions) is described in the NHS Five Year Forward View[i]*. This approach (summarised in Box 1 in the main report) is simplistic, and highly sensitive to anomalies in the underlying data.*

Among the weaknesses of this approach are:

1. *The inability to distinguish the noise from the true signal. The common approach is overly sensitive to high (or low) admission rates in a given stratum of the observed data. These anomalies will determine predictions (for that stratum) for all subsequent years.*
2. *The assumption that age-specific admission rates are constant over time. This is unlikely for two reasons. Firstly, it presupposes that the management of groups of patients – for instance the elderly and dying – will not change over time. Secondly, it fails to acknowledge*

[i] NHS Five Year Forward View, Recap briefing for the Health Select Committee on technical modelling scenarios. NHS England (2016)

any of the current theories of population ageing which propose different trends in the length of time that is spent in ill-health at the end of life.

3. *The absence of information regarding the statistical significance of model terms.*
4. *The high (model) degrees of freedom for the age variable compared to other modelling approaches.*
5. *The lack of versatility. One is confined to predicting admissions based solely on changes in the structure of the population.*

Alternative approaches include time series methods. These use trends in historical admission rates to forecast future activity and address or bypass many of the limitations of the traditional approach. There are a wide range of approaches of varying sophistication within this well-established field. However, we chose to address the issue of benchmarking activity levels by testing several types statistical regression model. Regression models generally provide predictive power, robustness and, importantly, versatility. In addition, they are often more interpretable than typical machine learning models and may also be blended with time series methods or other approaches.

The common approach and the (simple) GAM approach

We sought to compare the predictive capabilities of the common approach with a regression approach:

A typical regression model would assume the form:

$$\text{rate of admission} = B_0 + B_1 * (\text{age group X gender}) + \varepsilon, \quad \text{(Eq. 4)}$$

where B_0 and B_1 are the model parameters, and ε is a mean-zero random error term. We allow age and gender to interact.

There are a couple of additional points to consider in our case. The response variable, rate of admissions (count of admissions / population), is typically overdispersed. We therefore require a regression model which accommodates the negative binomial distribution.

Another important factor to consider in Eq. 4 is that the relationship between age and admission rate is highly non-linear (shown in Fig. 7, below).

These considerations led us to employ a generalised additive model (GAM) which supported a negative binomial distribution,

$$\text{rate of admission} = B_0 + f(\text{age group X gender}) + B_1 * \text{gender} + \varepsilon, \quad \text{(Eq. 1)}$$

where B_0 is the intercept, f is a smooth function of the age variable (interacting with gender), and ε is a mean-zero random error term. Note, this is not the same model used in the main report.

GAMs were built using the statistical software, R, version 3.5.1, and the package mgcv.

As noted, the relationship between age group and rate of admission is highly non-linear. Due to the severity of the gradient between the ages of 0 and 1, and 89 and 90+ (visible in Fig. 7), these age groups could not be appropriately modelled by the GAM without the risk of overfitting. As a compromise, we examined the admission rates for individuals aged between 1 and 89 in a principle model, and individuals aged 0 and 90 in two supplementary models.

Results

Rates of admission varied considerably by age group (Fig. 4). Age groups from 2 to 65 had less than 100 admissions per 1,000 population. However, children under 1 year have around 400 admissions per 1,000 population, while males over 90 years had 600 admissions per 1,000 population. Fig. 7 also highlights differences in rates due to gender. This is true mainly for the very young, for women of child-bearing age, and for the population over 70.

Fig. 8 shows the variation in admission rates by local authority for year 2013. The council area with the highest admission rates had twice the rates of the area with the lowest. There are a small number of council areas that have relatively high admission rates.

Figure 7. Admission rates, by age and gender, for the calendar year 2013.

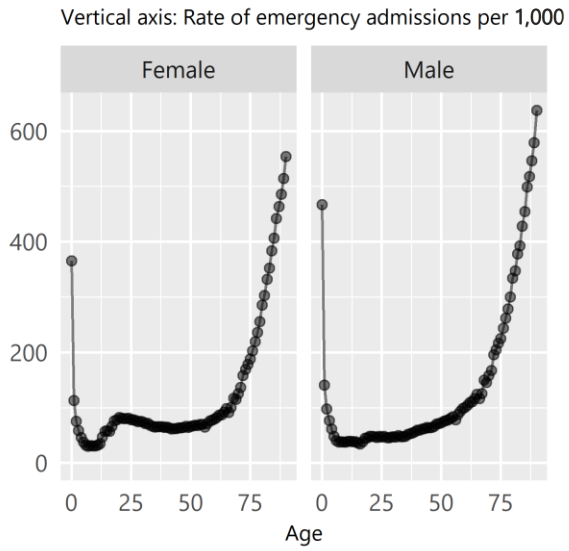
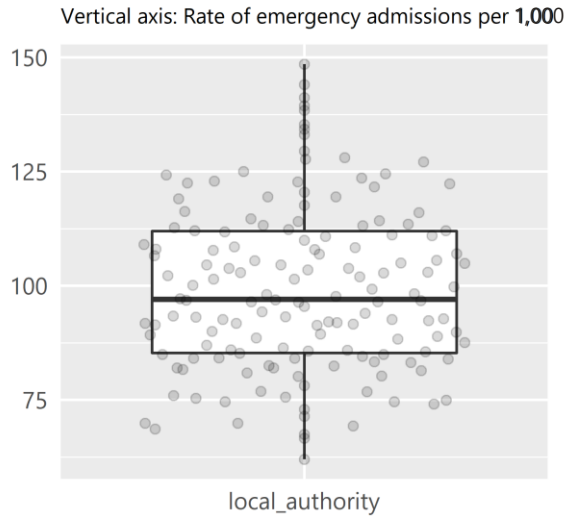


Figure 8. Admission rates, by local authority, for the calendar year 2013.



Predictions

Using data from 2013, we built a GAM for each of the 150 council areas. The median adjusted R squared for these models was 84.0 (78.0, 90.1).

With a validation dataset (consisting of three consecutive years; 2014 to 2016), we directly compared the accuracy of predictions from the GAMs with those returned by the common method. We chose the mean absolute error (MAE) for each council area model as a measure of model accuracy. Here, the residuals are the difference between the true and the predicted count of admissions, for each age and gender grouping within a council area.

Having computed 150 MAEs (one for each model) for each year in the validation dataset, we used the median MAE (MMAE) to measure of the overall performance of both approaches.

Principle Model (ages 1- 89)

For all years, the GAMs provided lower MMAEs than the common method (Table 2). Interestingly, the largest percentage difference in MMAE (21%) between the models was seen for a forecast one year in advance of the training data. The difference narrowed slightly as the forecast horizon grew more distant.

Table 1: Indicative accuracy (council area MMAE) of the common and GAM approaches for the principle model.

	Median MAE		
Model	**Forecast year 2014**	**Forecast year 2015**	**Forecast year 2016**
Common approach (principle)	22.35	23.34	25.37
Simple GAM (principle)	17.87 (- 21%)	19.05 (-18%)	21.10 (-17%)

Supplementary Models (age 0 and age 90)

The supplementary models (for ages 0 and 90) use single age groups, thus there is no need for smooth function for age. We would then be left with a simple negative binomial regression of the form:

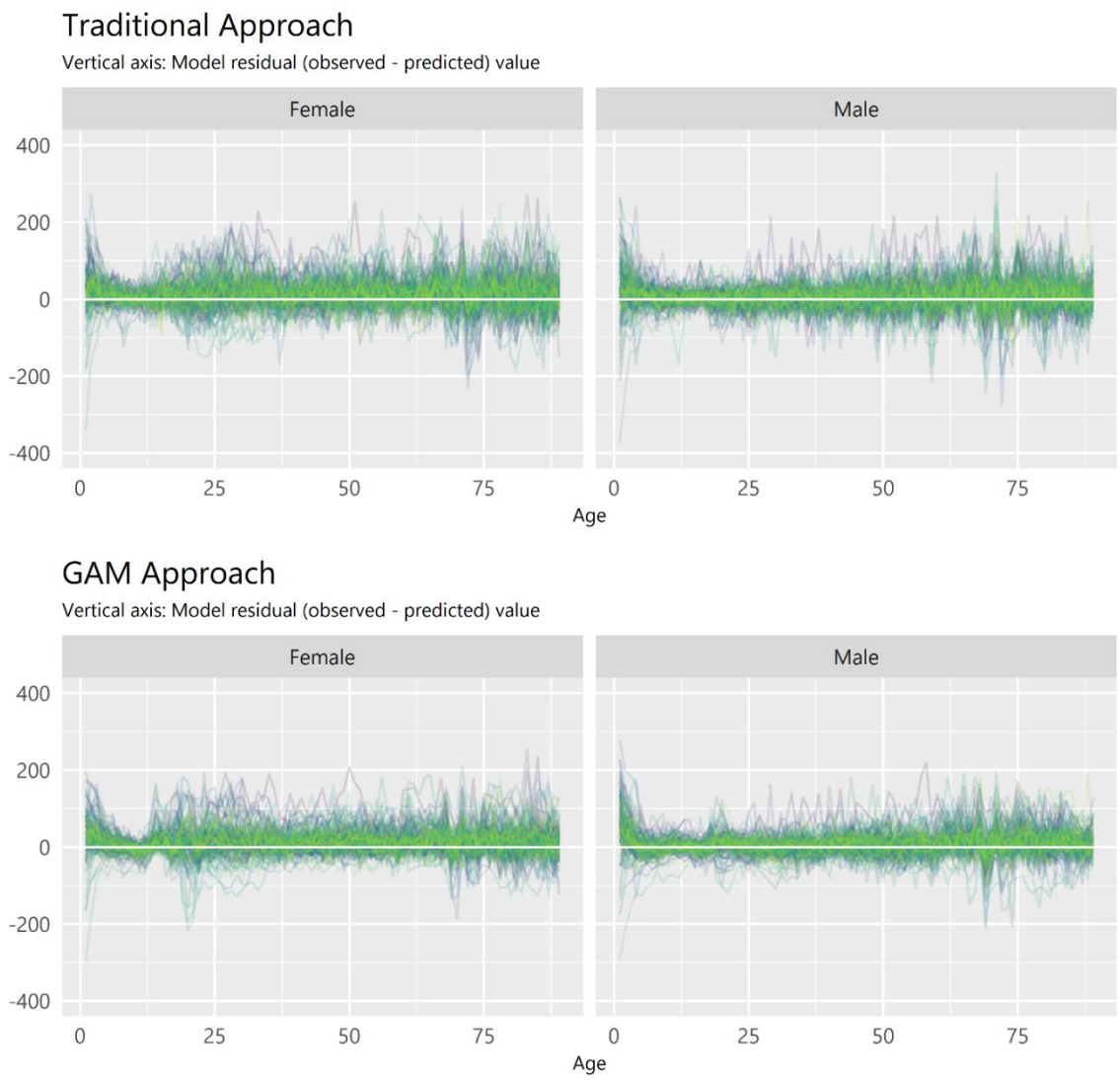
$$\text{rate of admission} = B_0 + B_1 * \text{gender} + \varepsilon, \quad \text{(Eq. 5)}$$

which, due to the fact that there are only two levels in gender, will perform the same calculation as the traditional method. Therefore, there is no difference between approaches here.

Comparison

We will therefore focus the a comparison of predictions on the principle models. Looking at a visual representation of model accuracy (Fig. 9) the majority of GAMs are tighter to the zero-residual line than for the traditional approach. The degree of symmetry about the zero-residual line indicates whether a model is over-estimating (negative residuals) or under estimating (positive residuals) the admission rate. Given this data for the year 2016, both approaches appear to broadly underestimate admission rates for both sexes, and the GAM for females displays the most asymmetry. With both approaches, there is high variation in residuals for the very young (those less than 3 years old), and for those aged 65 to 70. While the GAM shows less variation in deviance for those aged 65 to 70 years, it appears prone to overestimating the number of admissions for females around the age of 20. This last fact is true for all years (not shown).

Figure 9. A comparison of residuals from the common (top row) and GAM (bottom row) approaches, by gender and age group. Each line is a single council area and shows deviance of the model prediction from the actual value, by age. The forecast year is 2016 (three years beyond the training data).



Discussion

We have examined an alternative to the common approach used to forecast emergency admissions. The GAM approach appears to provide better prediction accuracy than the common approach. The mechanism behind this is the GAM's ability to estimate a smooth function which seeks to describe the non-linear relationship between age and admission rate. The GAM is therefore more robust to unusually high (or low) age-specific admission rates in the training data, caused by random (or non-random) variation in any given year.

While the appeal of the traditional method has always been its **simplicity**, *we highlight that the basic GAM employed in this analysis is just as easy to implement. Moreover, the GAM is versatile, it will provide us with While the appeal of the traditional method has always been its* **simplicity**, *we highlight that the basic GAM employed in this analysis is just as easy to implement. Moreover, the*

GAM is versatile, will provide us with statistical significance of model terms, and, unlike some of the more complex modelling approaches, we are able to see how the GAM arrives at predictions.

Limitations

There is a strong relationship between the quality of a model and the quality of the data on which the model is built. In studies which examine data collected over several years, and over broad geographical areas, there will be questions arising about the quality of data. Here, we must be particularly conscious of the consistency of coding practices. For example, it may be that the high variation in GAM predictions for females aged around 20, may be a result of inconsistencies in the coding of birth episodes.

When examining the modelling approaches themselves, we tested simple GAMs consisting of a few demographic predictors. These chosen variables undoubtedly have great influence on emergency admission rates. However, there are other important variables, both demographic (e.g. cohort) and non-demographic, which we did not investigate.

Looking at the demographic predictors that were included in the models, we were limited to the council area variable as a proxy for a measure of deprivation. This was due to a reliance on ONS population projections in the validation dataset. Thus, the model assumes that all individuals of a given age and gender within a council area will have same underlying health needs. While we do observe clear differences in admission rates between council areas (Fig. 8), in many - especially urban - areas there is great variation in the socio-economic status of individuals. Therefore, if lower-level geographic breakdowns of deprivation ranking could be incorporated into the models, this would very likely improve the forecasts.

In addition, the models assume that there are no changes in age-specific admission rates. This is unlikely for two reasons. Firstly, it presupposes that management of groups of patients, for instance the elderly and dying, will not change over time. Secondly, it fails acknowledge the current theories of population ageing which propose different trends in the length of time that is spent in ill-health at the end of life.

The Strategy Unit
Tel: 0121 612 1538
Email: strategy.unit@nhs.net
Web: www.strategyunitwm.nhs.uk
Twitter: @strategy_unit

NHS
Midlands and Lancashire
Commissioning Support Unit